Recent insights into the three-dimensional molecular packing structure of native type I collagen

Joseph P.R.O. Orgel[1,3], Andrew Miller[1,4], Thomas C. Irving[2], and Tim J. Wess[1].

[1] Center for Extracellular Matrix Biology, Department of Biological Science, University of Stirling, Stirling. FK9 4LA, Scotland UK.

[2] Center for Synchrotron Radiation Research and Instrumentation and the Dept. Biological, Chemical, and Physical Sciences, Illinois Institute of Technology, 3101 S. Dearborn St. Chicago, IL, 60616. USA.

[3] Current: Laboratory for Molecular Biology (MC 567), University of Illinois at Chicago, 900 S. Ashland Ave. Chicago, IL, 60607. USA.

[4] Current: Interim Chief Executive office, Imperial Cancer Research Fund, 61 Lincoln's Inn Fields, London WC2A 3PX. UK.

Received 28th December 2001; accepted in revised form 15th March 2002.

The structure formed by the packing of collagen molecules within fibrous tissue has been a matter of some considerable speculation and debate for decades. Its importance is due to the fact that the arrangement of collagen molecules within fibrils must certainly impart the basis of the strength and integrity of mammalian connective tissue, and determines how collagen fibrils interact with other extracellular matrix components.

We recently reported a solution for the structure of collagen type I in situ *[1]. This was obtained by modelling the x-ray fibre diagram and then solving the phase problem via MIR to produce our electron density map, thus avoiding any possible bias towards a particular structural model. In this paper, we review that work and present some additional details of the methods used that are particularly relevant to the fibre diffraction community.*

The diverse range of proteins within the collagen family form the major constituent of connective tissues in most vertebrates [2]. Collagen is so prevalent that it is believed to be the single most abundant protein in the animal kingdom, accounting for approximately one third of the total protein mass in animals. Amongst these, collagen type I is the predominant form. Yet progress in understanding the organization of mammalian connective tissue at a basic level (the molecular arrangement) has been disproportionately slow when related to collagen's abundance, strategic functions and locations, and relative to the dizzying pace of structural solutions for soluble proteins through crystallography and NMR. A major factor impeding the progress of collagen structure research has been the uncertainty inherent in the previous approach of comparing the calculated diffraction pattern of various stochastic models of the molecular packing with the observed diffraction pattern. Such an approach may permit discrimination among the chosen models but it cannot of itself provide assurance that a more accurate model, yet to be identified, does not exist. We have attempted to avoid possible bias by restricting the model generation process to that of the diffraction pattern alone, and then calculating phases from amplitudes obtained from native and isomorphously labelled samples. Obtaining the intensities from the diffraction pattern using this modelling approach and then obtaining the structure through MIR was made possible by a culmination of efforts, some of these efforts having lasted through the better part of one of our author's careers.

We present here a review of this challenging structural problem, and how we resolved it using a combination of fibre diffraction and x-ray crystallographic methods.

Collagen structure

The molecular organization of a single fibril is schematized in Figure 1. When heavily stained, collagen fibrils seen in electron micrographs exhibit the well-known dark/light-banding pattern, caused by the deposition of heavy metal into the gap resulting from a shorter fifth molecular segment within the five molecular segment unit cell. This axial structure is well defined. Meridional reflections in the fibre diagram extend beyond 0.45 nm (over

150 reflections). The lateral packing (molecular packing arrangement in cross-section of the fibril) was until very recently much less well characterized, except for the understanding that collagen molecules are somehow arranged on a quasi-hexagonal lattice [3,4]. It is the characterization of this part of the structure that has been so elusive.

The collagen x-ray fibre diagram, a technical challenge

A strong, diffuse 'fan' underlies the off-meridinal parts of the collagen fibre diagram (Figure 2). Whilst containing significant information in its own right, it is treated as 'background' in the attempt to quantify the Bragg scatter appearing in the same regions of reciprocal space. Its presence suggests that there is significant disorder within the tendon structure, with the maximum intensity being in the region of the intermolecular interference function. A later study [5] concluded that the diffuse scatter contains a significant contribution from the gap region where there is a lower packing density due to the missing molecular segment. Consequently, the remaining four segments have a greater tendency to motion [6] and/or reduced regularity in their molecular positioning. The molecules within the relatively well ordered overlap region have greater packing density (five chains) and the intermolecular crosslinks at the gap/overlap regions probably have much less freedom for azimuthal motion and possess greater positional regularity than the gap region, which shows more extensive disorder. The consequences for the x-ray fibre diagram are that the contribution of the gap region to higher angle off-meridonal reflections is attenuated. Despite this, the three-dimensional x-ray fibre diagram was successfully indexed on a triclinic unit cell after the strong near-equatorial reflection at 1.3 nm (a spacing close to that expected from near neighbor molecules), was shown to be split into three strong components and suggested a quasi-hexagonal packing of collagen molecules [3]. An improved set of unit cell parameters was later deduced by more detailed measurements [7,8].

A further complication to the interpretation of fibre diffraction diagrams in general and in the case of those arising from tendons, is that they are not single crystals, but are made up from millions of fibrils that are probably not all themselves crystalline since only 10% of the total off-meridional intensity arises from Bragg scatter [9]. Even here, since the collagen fibrils are not truly aligned with a spread of around 1-2 degrees from true parallelism, the Bragg reflections in the fibre diagram are often drawn out into arcs that overlap with other reflections. These two features present enough technical challenges, but in addition to these complications, the crystallographic unit cell is triclinic (p1) with its

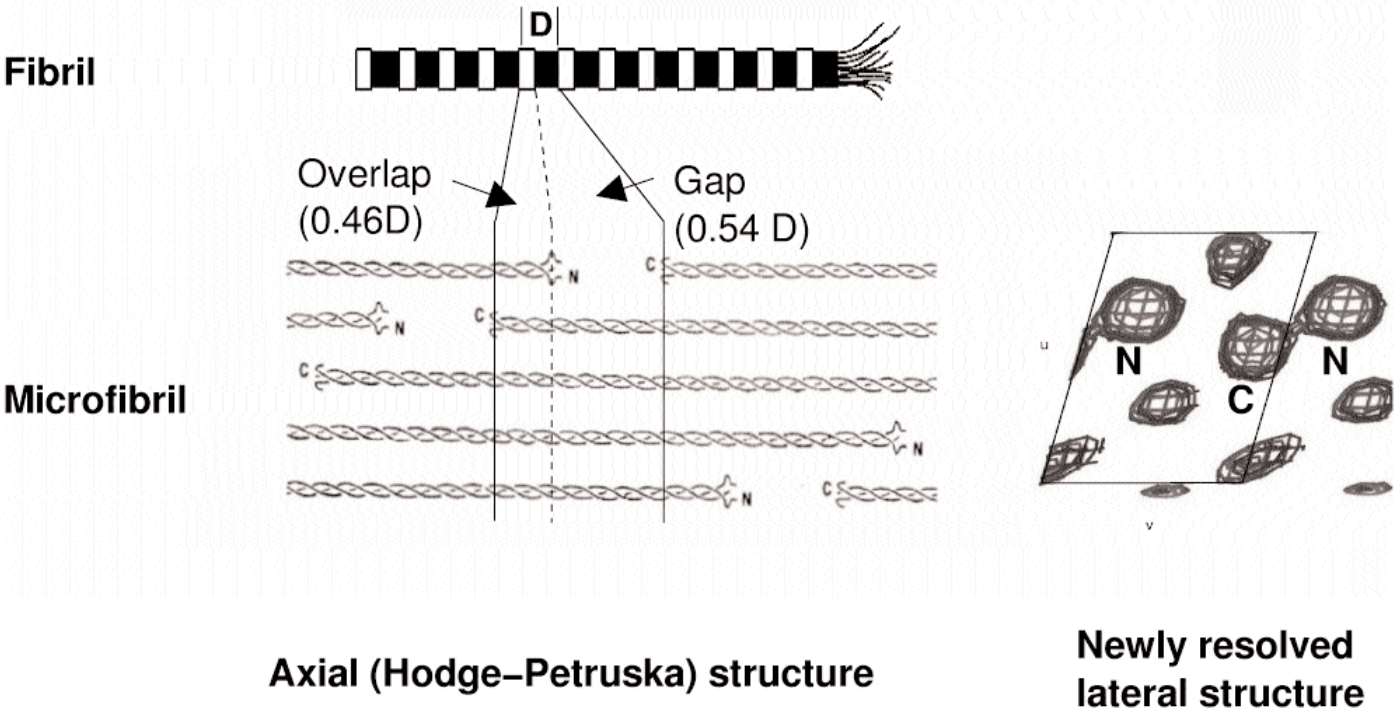

Figure 1: Axial and lateral molecular composition of the type I collagen fibril. N and C termini are labeled. The axial unit cell (D) is shown (left), as is an outline of the unit cell of the molecular packing structure (right). Molecular segments from a neighboring unit cell are shown (right) to better show the quasi-hexagonal nature of the packing structure and the corrugated pattern of the crosslinked N and C telopeptide containing molecular segments within the packing scheme.

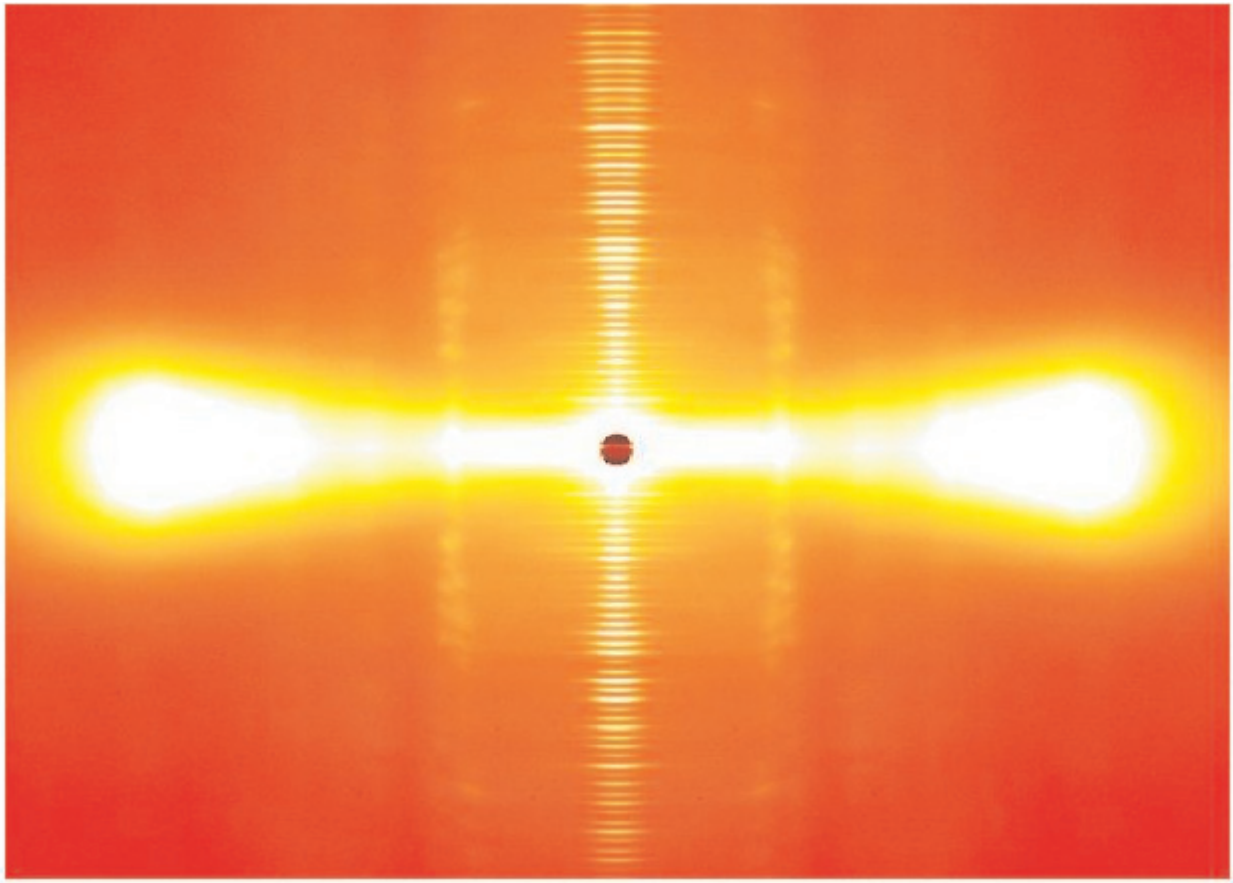

Figure 2: X-ray fibre diagram of rat-tail tendon, iodide derivative before data treatment.

(very) long axis (c) inclined at 5 degrees to the axis of the fibril. Finally, since the diffracting specimen is a fibre, the x-ray pattern is equivalent to a 360-degree rotational pattern from a single crystal.

Tackling the challenge

A great deal of improvement in the quality of x-ray diffraction patterns from collagen specimens has been made in recent years. The use of synchrotron radiation and new image treatment techniques have allowed the measurement of the intensities of the *(h,k,l)* reflections from native tissue and from isomorphous derivatives treated with iodine (which labels the tyrosine amino acids which only occur in the telopeptides) and gold which labels histidine, selected methionine and cystine residues (of which only the former two are found within the sequence of type I collagen). In these latter cases, it has been possible to improve markedly the visibility of weak Bragg reflections, that may be only barely visible or not seen at all, in the native pattern, but still enable quantitative interpretation of the data since these heavy atom treatments produce predictable staining patterns within a relatively small number of residue attachment sites [10,11].

Even so, several of the problems discussed above had not, until recently, been dealt with. These were:

a) Obtaining data of superior quality to allow the separation of the closely spaced Bragg reflections, (especially in the near-equatorial region) from the diffuse scatter.

b) The development and application of techniques to separate and determine the intensity of reflections that overlap in the off-meridonal parts of the fibre diagram.

c) Repeating this process for two or more isomorphous derivative proteins, the data from which would then be used to determine the three-dimensional location of binding sites, and eventually the native phases.

Improving data quality

Several preliminary experiments were performed at beamline 7.2, Synchrotron Radiation Source (SRS) Daresbury UK, and beamline ID2, European Synchrotron Radiation Facility (ESRF) Grenoble, France, in the process of developing stain techniques and in parallel with the determination of a high angle one-dimensional structure of collagen [10]. Eventually, we were able to obtain data sets of

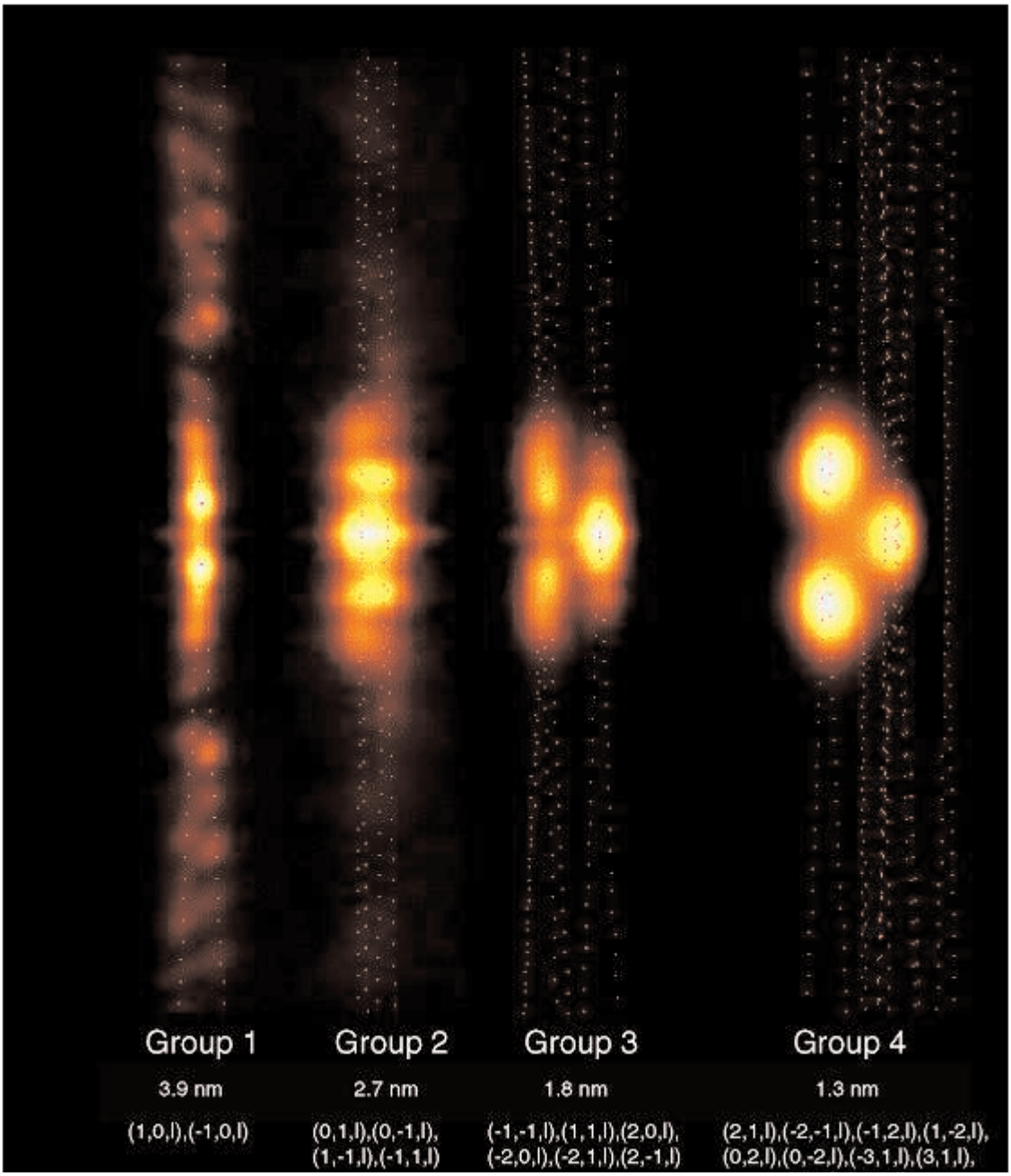

Figure 3: Portion of the fibre diagram showing Figure 2 after background subtraction. Bragg positions are marked and the groups 1-4 (with hkl reflections and reciprocal space ranges listed) are also labelled. Adapted from Orgel et al., 2001.

sufficiently high quality during an extended experiment at the BioCAT ID 18 beamline, Advanced Photon Source (APS), Argonne Illinois USA. In these fibre diagrams, the low beam divergence, small beam size, and clean optics of this instrument [12] have greatly reduced the parasitic scatter and the associated background and achieved a high order to order resolution (Figure 2, and Figure 3). The diffuse scatter of the pattern due to the inherent disorder contained within the samples remains, and was dealt with as described below.

A camera length of 1026 mm (at λ = 1.003 Å) was chosen to maximize the number of pixels over which fibre diagram extended (i.e. improve the resolution in Q space), whilst maintaining the presence of the reflections at higher Q-space values needed solve the phase problem within the same diffraction pattern. This optimization would later prove critical during data extraction (see Table 1). The removal of the remaining 'background' (diffuse scatter), was performed by methods used previously [8,9], but worthy of brief review here:

The suite of programs FIT2D [13] was used to remove the diffuse scatter from the low-angle diffraction pattern background. This background was assumed to be smooth and regions containing background were presumed to be distinguishable from regions that did not. One of the drawbacks of this assumption is that in areas containing Bragg

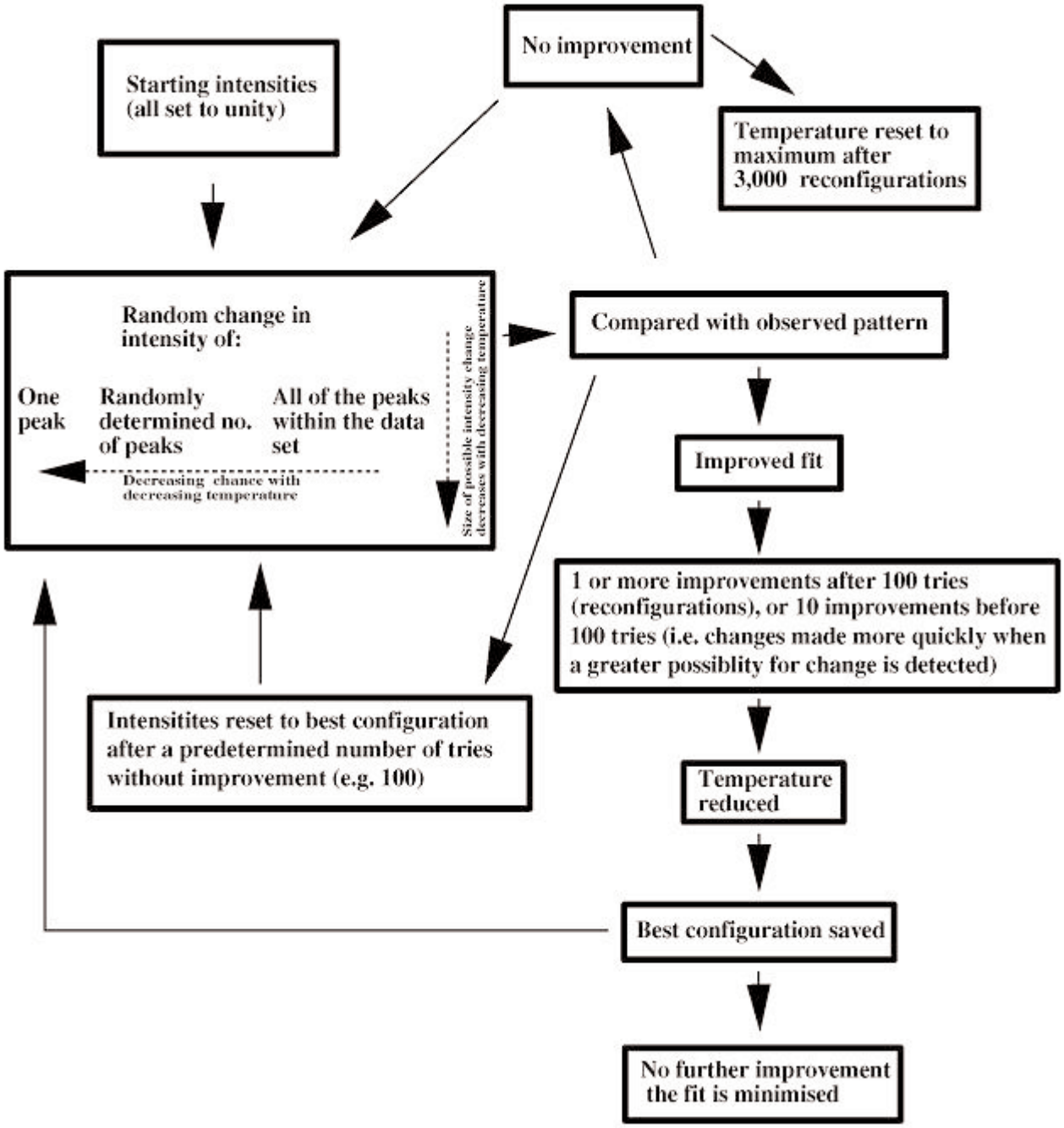

Figure 4: Schematic showing the operation of the simulated annealing algorithm that was used to determine the intensities of the native and two isomorphous derivative samples.

scatter that happens to be indistinguishable from the underlying background, the smooth background polynomial would model only a background function. Hence any Bragg scatter (the sought after data) in such areas would be partially or completely removed in the background subtraction process. The likely-hood is that weaker Bragg intensity reflections will be effected (reduced or removed) in this process more than stronger reflections will be. However, the diminished influence of the weaker reflections in the phase calculation is preferable to the over-representation of reflections that happen to be located on top of a highly sloping background that would otherwise adversely effect the phase and electron density calculations if left uncorrected.

Interpolation was performed within FIT2D in a user-defined area of the pattern (between approximately R= 0.15 to 1.0 nm^{-1} and Z= 0.0 to 0.4 nm^{-1}), with pixels containing Bragg peak data masked (excluded from the calculation). This area of the diffraction pattern contains the main part of the small-angle off-meridional data. Surface polynomial functions (two-dimensional Chebyshev polynomials) were used to fit a background model to the data. Generally, the area fitted would be larger in the Y direction than the X, hence the X polynomial would be in the order of 3-5 whilst the Y would be approximately 5-10. In either case, the polynomial order would always be much smaller than the number of pixels within the

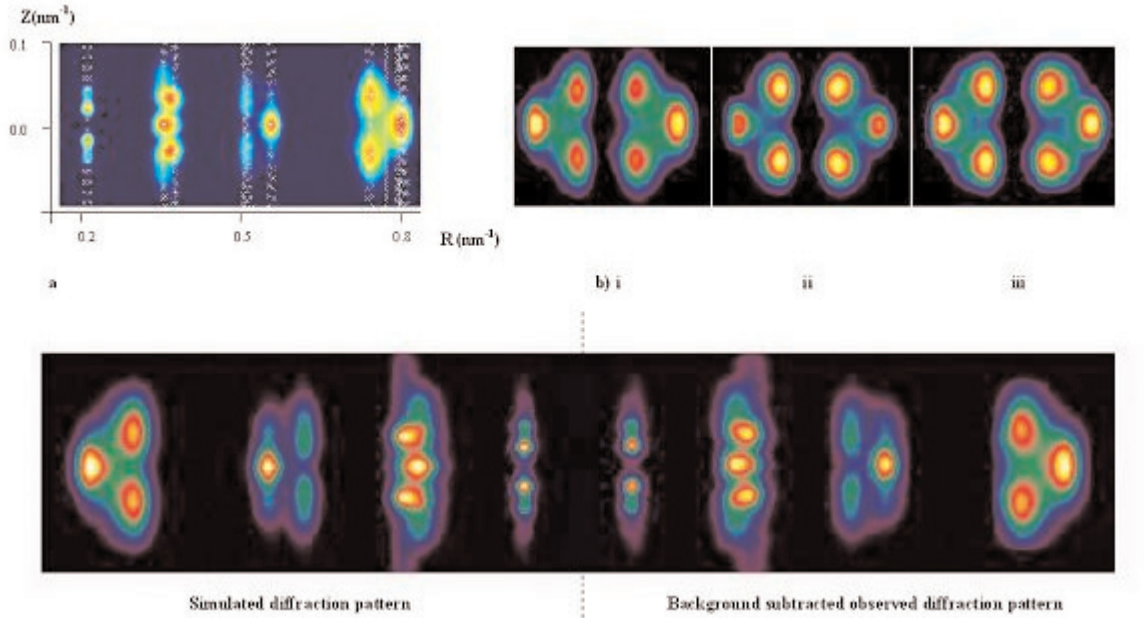

Figure 5: Comparisons of background-subtracted portions of the observed fibre diagrams with the computer generated patterns used in the determination of the overlapping reflections

a) A background subtracted low angle off-meridional region of the native fibre diagram, the position of the Bragg reflections are marked with X's.

b) Group 4 rowlines that contain the highest degree of Bragg overlap within the fibre diagram. This shows the high degree of agreement between the observed and the computer generated patterns; the observed is on the right, the simulated on the left of each of the following:
i. Native, ii. Iodide derivative, iii. Gold chloride derivative.

c) Right, The observed background subtracted image shown in a). Left, the corresponding simulated low angle diffraction pattern. This is the region containing the highest degree of Bragg reflection overlap, and, for the native data set that shows the highest percentage error between that of the observed and simulated patterns (3.00%). Adapted from Orgel et al., 2001.

Figure 6: Tilted molecular segments of the overlap region

The molecular tilt of the collagen molecular segments of the overlap region are observed to follow a vector approximately parallel to the line (0,0,0) to (0,2,1) (u,v,w; as in Fraser et al. [4]). This corresponds to a tilt of about 5 degrees relative to the c-axis of the unit cell. The molecular segments in the overlap region follow parallel paths. The formation of intermolecular crosslinks at the interfaces of the overlap/gap regions (at the telopeptides), ensures that the overlap region is well ordered, particularly in the plane of the telopeptides in contrast to the disordered state of the gap [5] see also [16]. The overlap region is illustrated here and the c-axis compressed by 5 times to show clearly the tilt of the individual molecular segments. The electron density is the same as that seen in Orgel et al., 2001, although the view is similar, the sigma levels (density scale) are slightly lower to contrast gap and overlap regions.

interpolation region.

We assumed that we had achieved a suitable background model when the residual between the model and observed data was minimal [13]. We used the lowest order polynomial that fitted these criteria, so that the rowlines were seen to be superimposed on a suitably flat background. This process was made considerably easier because of the improved quality of the fibre diagrams.

Extracting the data

Wess *et al*., [9] had previously developed a model for fitting the group 1 row-lines (where group n (n is an integer 1-4) refers to the groups of row-lines in the collagen fibre diagram, 1 being closest to the meridian, 4 being furthest away, see Figure 3) using a two dimensional Gaussian function for each peak, the coordinate positions of each peak being calculated from the unit cell parameters of Wess *et al.*, [8]. We adopted and further developed this strategy using a computer generated diffraction pattern to fit the background-subtracted observed diffraction pattern, the minimisation between the two being driven by a simulated annealing method developed from the original concept described by Metropolis *et al*., [14].

An algorithm was written following these principles, applying the annealing analogy to fitting a model diffraction pattern to the observed diffraction pattern, (the algorithm is summarized in the schematic flow diagram of Figure 4), each peak of intensity being treated as an atomic position in the annealing analogy. The coordinate positions of the intensities in the diffraction pattern were fixed according to the unit cell parameters previously described [8].

Three kinds of intensity change (analogous to atom redistribution in the annealing analogy) were allowed for each peak of the model pattern: 1) A random intensity change between the minimum and maximum pixel values of the observed pattern, of one randomly determined peak. 2) One random intensity change for a selected sequential series of peaks of random length. 3) One random global intensity change for all the peaks being fitted.

The model system had an imposed temperature, just as in the thermodynamic analogy. At high temperature, greater intensity changes were allowed as well as more wide spread change to the number of peaks affected. As the temperature dropped, the chances of the options 2 or 3 (above) occurring diminished, reflecting the fact that, as in the thermodynamic analogy, widespread changes affecting arrays of atoms would be less likely to occur as the liquid cooled.

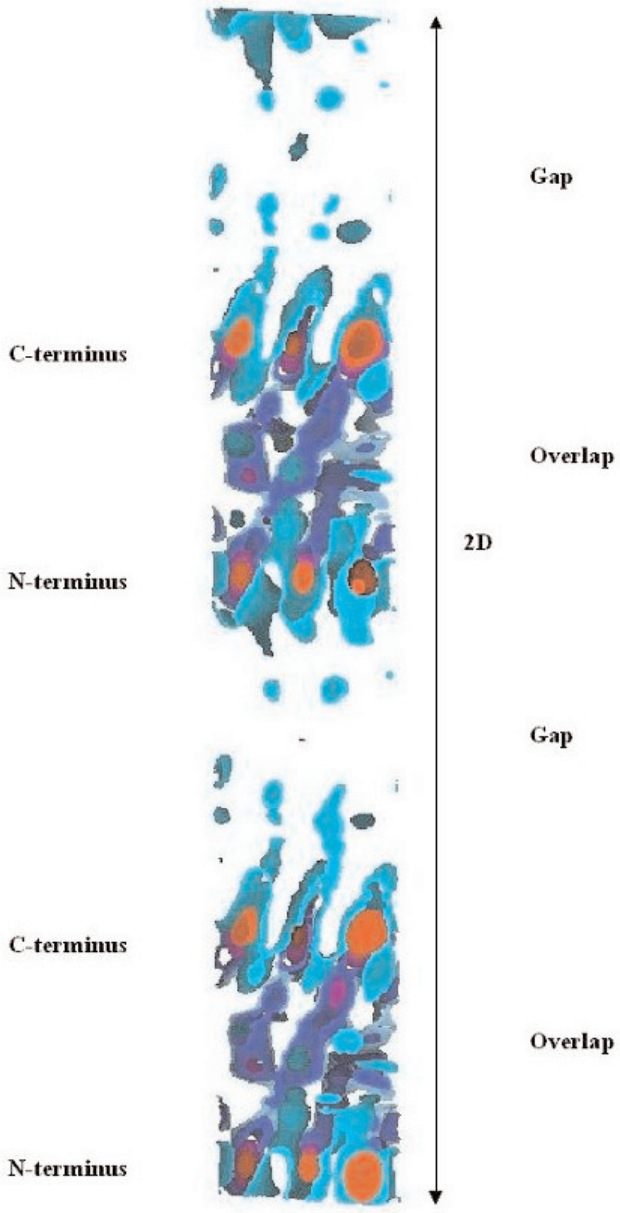

Figure 7: Two unit cells (along long unit cell axis) showing ordered overlap and disordered gap regions.

The molecular packing arrangement is most clearly discernible at the axial level of the telopeptides, (shown as the C terminus and N terminus regions on the figure) but is also discernible within the rest of the overlap region. The gap region shows a large degree of lateral disorder of the molecules in the lower packing density gap region, as predicted in several model studies [5,7,16]. Disorder is commonly encountered in macromolecular crystallography in regions of crystal structures subject to thermal motion (such as chain loops), and was seen to a similar degree in the study of tropomyosin [18]. The electron density is the same as that seen in Orgel et al., 2001, although the view is similar the sigma levels (density scale) are slightly lower to contrast gap and overlap regions.

Source	ID18 BioCAT undulator beamline(APS)
Beamsize (unfocused)	0.8 mm x 3.6 mm (at sample) [12]
Flux	1.5-5 x 10^{13} ph/s @ 100 mA
Pathlength	1026mm
Wavelength	1.003 Å
Detector	Fuji BAS V image plates with a Fuji BAS2500 reader
Data sets	One native and two isomorphous derivative fibre diagrams
Derivatives	iodide and gold chloride
Unit cell dimensions	triclinic a=39.97 b=26.95 c=677.9 Å
Unit cell angles	α=89.24º β=94.59º γ=105.58º
Space group	P1
Resolution limits (Å)	5.4 meridional 10.0 equatorial
Observed unique reflections	410 (124 meridional 286 equatorial)
Figure of merit	0.4

Table 1: Experimental details. Adapted from Orgel *et al*., 2001.

Data set	Group	1	2	3	4
	Average error (%)				
Native	3.02	3.23	4.21	1.57	2.99
Iodide	1.56	0.97	1.69	1.48	2.08
Gold	1.34	1.64	1.55	0.64	1.52

Table 2: Percentage error between simulated patterns and the observed fibre diagrams. Adapted from Orgel, *et al*., 2001.

The algorithm would run until it was assessed that no further minimisation would occur. The success of each cycle and the final minimisation was judged through the use of the Root-mean-square difference in each pixel between the observation and simulation. Only the essential parts of the fibre diagram were calculated during the minimisation these being: all of the Bragg reflections, their Gaussain function and a small amount of the flat background surrounding each peak (around 0.09% per reflection of the total data area). This was to improve the speed of the algorithm only requiring it to make 1/100 of the calculations it would otherwise have needed to make if it were calculating the whole of each fibre diagram (most of which had already been assessed as flat after background subtraction of the observed data, described above). The R-factor was calculated to assess the fit of each simulated pattern to its respective observed pattern as follows:

$$R = \Sigma_i \mid F_o(i) - F_c(i) \mid \; / \; \Sigma_i \; F_o(i)$$

Where F_o and F_c are respectively the observed and calculated intensity of each pixel.

The final combined error for the native and the two derivative patterns was judged to be less than 3.0%. (Figure 5 and Table 2). This low level of error gave us sufficient confidence in the intensities so derived, to proceed with the phasing process. The process of phasing would test the real success of these methods; too much error in the assessment of the intensities could make it impossible to determine the phases accurately enough to be able to visualize the molecules and elucidate the packing system.

Determining the phases

As in many crystallography experiments, the determination of the structure factor phases can be the hardest and most laborious part of the data processing, especially when the sites of heavy atom

attachment are not known. Fortunately, data provided from the high angle one-dimensional structure of Orgel *et al.*, [10], identified the putative residues, their axial location, and the molecular segments within which they were bound. This greatly reduced the burden of heavy atom searching to that of looking for heavy atom peaks in difference Patterson maps corresponding to the axial distance between the telopeptide regions (the major sites of heavy atom attachment). The two derivatives, gold chloride and iodide treated samples each possessed two major sites for heavy atom attachment. These were clusters of residues that accept heavy atom stain, located, for the gold chloride derivative, at the N-terminus within molecular segment 5 and at the C-terminus within molecular segment 1, and for iodide derivative: at the N-terminus within molecular segment 1, and at the C-terminus within molecular segment 5 (see reference 10). Difference Patterson maps were calculated for the packing structure at the axial separation of the telopeptide regions (N and C-termini) of 0.33-0.41 w (fractional unit cell coordinates). The resulting vectors for the gold chloride and iodide derivatives were respectively: u=0.2, v=0.4 and u=0.05, v=0.25 (where u and v are fractional unit cell coordinates in Patterson space).

These values and the corresponding Patterson maps were used in conjunction with autocorrelation functions (calculated using XtalView programs [15]) to locate the three-dimensional positions of the molecular segments 1 and 5. When agreement was reached, these coordinates were used to calculate phases that were refined using the Xheavy program of the Scripps Institute crystallographic software suite XtalView [15].

The first true x-ray diffraction visualisation of a semi-crystalline natural fibre

The end result is shown in part in Figures 6 and 7. The molecular segments within the overlap region are clearly visualized (particularly at the telopeptide regions), whereas little is seen of them within the gap region. In cross section, the quasi-hexagonal packing scheme is revealed, (with the intermolecular cross-linking telopeptides identified in the difference Fourier maps, see [1]). The crosslinking pattern and its significance to the supermoleculer structure is discussed elsewhere [1], but it is significant to see that the molecular segments of the overlap region are tilted with respect to the unit cell axis, as predicted by Fraser *et al.*, [7]. This is testament to the successful extraction of the data from an extremely difficult system.

The successful generation of an interpretable electron density map that fits the consensus elements of several model studies [3-5,7-9] is compelling proof of the validity of the methods used to extract the intensities. It is still important, however, to continue to improve the quality of the electron density map. Work has already commenced to attempt to obtain improved diffraction patterns from fibrils that are cryofrozen. The freezing of the tendons in this way will allow the sample to be bathed in the x-ray beam for longer to obtain a higher signal to noise ratio, and further reduce the thermal disorder of the molecular system within the sample. It may be possible to reduce the disorder of the fibrillar packing arrangement using a series of freezing/partial thawing steps to anneal the packing system to a lower energy state.

An altogether different approach, still using x-ray fibre diffraction techniques, involves attempts to obtain diffraction from a *single* fibril. This approach would be closer to true single crystal macromolecular crystallography, and would, if it can be made to work, produce diffraction patterns with improved spatial resolution of the Bragg peaks (reduced range of fibril orientations within the fibre - only one fibril). Initial experiments (at beamlines ID 22 and 18F at the ESRF, Grenoble, France) have shown the necessity of cryofreezing the sample whilst using powerful x-ray beams of very small cross-sectional area (approximately 0.75 μm^2). As the number of unit cells sampled by the incident x-ray beams is diminished in samples of this size (micron scale), the exposure time needed to collect data is increased (around 10-100 fold for this particular system). The cumulative effect of focused, high-brilliance synchrotron beam and tiny samples is the rapid destruction of delicate specimens, and inspiration will need to be drawn from the related field of macromolecular crystallography for techniques that preserve the life-span of samples over the extended time needed to collect data.

Both of these techniques involve treating the collagen sample in ways that are non-physiological. It is possible that the freezing of samples in the annealing studies, or the disruption of the sample to obtain single fibrils, would fundamentally change the molecular packing arrangement within the fibril. Solutions devised from these and other means,

however, can now be compared directly to the visualisation of the hydrated native protein presented here. This represents the first structure determined by MIR for any natural fibre, and the first three dimensional visualisation of the three-dimensional, periodic lateral packing structure of collagen.

Correspondence and requests for materials should be addressed to J.Orgel
(e-mail: jorgel@tigger.cc.uic.edu).

References

[1] Orgel, J.P.R.O. Miller, A. Irving, T.C. Fischetti, R.F. Hammersley, A.P. and Wess, T.J. The *in situ* three dimensional packing structure of type I collagen". *Structure*, **9**, 1061-1069. (2001).

[2] Hulmes, D.J., The collagen superfamily - diverse structures and assemblies. *Essays Bichem*. **27**, 49-67. (1992).

[3] Hulmes, D. J. S., & Miller, A. Quasi-hexagonal molecular packing in collagen fibrils. *Nature*. **282**, 878-880 (1979).

[4] Fraser. R. D. B., Macrae. T. B. & Miller. A. Molecular packing in type I collagen fibrils. *J. Mol. Biol.* **193**, 115-125 (1987).

[5] Hulmes, D.J.S., Wess, T. J., Prockop, D. J., & Fratzl, P. Radial packing, order and disorder in collagen fibrils. *Biophys. J.* **68**, 1661-1670 (1995).

[6] Torchia D.A., Solid state NMR studies of molecular motion in collagen fibrils. In *Methods of Enzymology*, **82**. 174-186. Cunningham, L., and Fredriksen, D., eds. Academic Press, New York (1982).

[7] Fraser, R. D. B., MacRae, T. P., Miller, A., & Suzuki, E. Molecular conformation and packing in collagen fibrils. *J. Mol. Biol.* **167**, 497-521 (1983).

[8] Wess, T. J., Hammersley A., Wess L., & Miller A. Type I collagen packing, conformation of the triclinic unit cell. *J. Mol. Biol*. **248**, 487-493 (1995).

[9] Wess, T. J., Hammersley A., Wess L., & Miller A. Molecular packing of type I collagen in tendon. *J. Mol. Biol.* **275**, 255-267 (1998).

[10] Orgel, J.P., Wess, T.J., & Miller, A. The in situ conformation and axial location of the intermolecular cross-linked non-helical telopeptides of type I collagen. *Structure*. **8**, 137-142 (2000).

[11] Bradshaw J.P., Miller A., & Wess T.J. Phasing the meridional diffraction pattern of type I collagen using isomorphous derivatives. *J. Mol. Biol.* **205**, 685-694 (1989).

[12] T.C. Irving, R. Fischetti, G. Rosenbaum, and G. B. Bunker (2000). Fiber Diffraction Using the BioCAT Undulator Beamline at the Advanced Photon Source. *Nuclear Instruments and Methods(A)* **448**, 250-254.

[13] Hammersley, A. P. FIT2D V9.129 Reference Manual V3.1. Grenoble: European Synchrotron Radiation Facility, 1998, www.esrf.fr (or through http://biocat1.iit.edu/FIT2D/).
[14]. Metropolis N., Rosenbluth M., Teller A., & Teller E. Equation-of-state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087 (1953).

[15] McRee D.E., *Practical Protein Crystallography*, Academic Press, San Diego CA. (1993).

[16] Bailey A.J., Light N.D., & Atkins E.D.T. Chemical cross-linking restrictions models for the molecular organization of the collagen fibre. *Nature*. **288**, 408-410 (1980)

[17] Woodhead-Galloway, J., & Machin, P., Modern theories of liquids and the diffuse equatorial x-ray scattering from collagen. *Acta Cryst. A.* **32**, 368-372 (1976).

[18] Phillips G.N., Fillers J.P., and Cohen C. Motions of tropomyosin. Crystals as metaphor. *Biophys. J.* **32**, 485-502 (1980).